

# KCRC-LCD: Discriminative Kernel Collaborative Representation with Locality Constrained Dictionary for Visual Categorization

Weiyang Liu<sup>a</sup>, Zhiding Yu<sup>b,\*</sup>, Lijia Lu<sup>a</sup>, Yandong Wen<sup>c</sup>, Hui Li<sup>a</sup>, Yuexian Zou<sup>a</sup>

<sup>a</sup>*School of Electronic and Computer Engineering, Peking University, China*

<sup>b</sup>*Department of Electrical and Computer Engineering, Carnegie Mellon University, U.S.*

<sup>c</sup>*School of Electronic and Information Engineering, South China University of Technology, China*

## Abstract

We consider the image classification problem via kernel collaborative representation classification with locality constrained dictionary (KCRC-LCD). Specifically, we propose a kernel collaborative representation classification (KCRC) approach in which kernel method is used to improve the discrimination ability of collaborative representation classification (CRC). We then measure the similarities between the query and atoms in the global dictionary in order to construct a locality constrained dictionary (LCD) for KCRC. In addition, we discuss several similarity measure approaches in LCD and further present a simple yet effective unified similarity measure whose superiority is validated in experiments. There are several appealing aspects associated with LCD. First, LCD can be nicely incorporated under the framework of KCRC. The LCD similarity measure can be kernelized under KCRC, which theoretically links CRC and LCD under the kernel method. Second, KCRC-LCD becomes more scalable to both the training set size and the feature dimension. Example shows that KCRC is able to perfectly classify data with certain distribution, while conventional CRC fails completely. Comprehensive experiments on many public datasets also show that KCRC-LCD is a robust discriminative classifier with both excellent performance and good scalability, being comparable or outperforming many other state-of-the-art approaches.

**Keywords:**

Kernel Collaborative Representation, Regularized Least Square Algorithm, Nearest Neighbor, Locality Constrained Dictionary

## 1. Introduction

Recent years have witnessed the great success of the sparse representation techniques in a variety of problems in computer vision, including image restoration [1], image denoising [2] as well as image classification [3, 4, 5]. Sparse representation is widely believed to bring many benefits to classification problems in terms of robustness and discriminativeness. Specifically, sparsity is a regularizer that can reduce the solution space under ill-conditioned problems, by seeking to represent a signal as a linear combination of only a few bases. These bases are called the “atoms” and the whole overcomplete collection of atoms together form what one call a “dictionary”. Many natural signals such as image and audio indeed have sparse priors. Imposing sparsity not only returns a unique solution, but also helps to recover the true signal structure, giving more robust estimation against noise. In addition, the sparse representation of a signal often leads to better separation and decorrelation which benefits subsequent classification problems. Despite the fact that sparse optimization is a nonconvex problem, the  $l_1$ -norm convex relaxation and its optimization techniques have been thoroughly studied.

Wright et al. [3] employed the entire set of the training samples as the dictionary and reported a discriminative sparse

representation-based classification (SRC) with promising performance on face recognition. SRC approximates an input signal  $\mathbf{y}$  with a linear combination of the atoms from an overcomplete dictionary  $\mathbf{D}$  under the sparsity constraint and gives the predicted label by selecting the minimum reconstruction residuals. Despite the fact that SRC was widely used in various applications [6, 7, 8], [9, 10, 11] still questioned the role that sparse representation plays in the image classification tasks.

Zhang et al. [11] further commented that it is unnecessary to enforce the sparse constraint with computationally expensive  $l_1$ -norm if the feature dimension is high enough. Their work emphasized the importance of collaborative representation (CR) rather than the sparse representation, arguing that CR is the key to the improvement of classification accuracy, which was validated by their comparison experiments. They used the  $l_2$ -norm regularization instead of  $l_1$ -norm, further improving the classification accuracy while significantly reducing much computational cost. The corresponding proposed method is called collaborative representation classification (CRC).

Despite their robust performance, the linear nature of both SRC and CRC makes them perform poorly when the training data are distributed vector-like in one direction. Kernel function, proven useful in kernel principle component analysis (KPCA) [12] and support vector machine (SVM) [13], was introduced to overcome such shortcoming for both SRC and CRC, leading to the kernel sparse representation-based classifi-

\*Corresponding author

Email address: yzhiding@andrew.cmu.edu (Zhiding Yu)

cation (KSRC) [14] and the kernel collaborative representation classification (KCRC) [15]. In particular, a Mercer kernel implicitly defines a nonlinear mapping to map the data from the input space into a high or even infinite dimensional kernel space where features of the same class can be grouped together more easily and different classes become linear separable.

Besides summarizing the KCRC [15], our major contribution in this paper lies in proposing a generalized framework for KCRC with locality constrained dictionary and unified similarity measure, giving both performance gain and significant reduction of computational cost. Due to the poor scalability of the global dictionary (GD) used in CR-based methods, classification becomes intractable in large database for KCRC with GD (KCRC-GD). To enable the scalability to large databases, we prune the dictionary via  $k$ -nearest neighbor (K-NN) classifier to enforce locality. Specifically, the nearest neighbors of a query serve to construct a locality constrained dictionary (LCD) for KCRC. Such strategy is both intuitively reasonable and mathematically appealing. Intuitively, LCD is well motivated by the psychological findings about human perception that visual categories are not defined by lists of features, but rather by similarity to prototypes [16]. In other word, coarse level matching, for which K-NN is used, plays an important role in human perception. Mathematically, LCD can be nicely incorporated under the framework of KCRC. First, the LCD similarity measure can also be kernelized under KCRC, which theoretically links CRC and LCD under the kernel method. Second, KCRC-LCD becomes more scalable to both the training set size and the feature dimension. The kernel gram matrix is now obtained from a subset of the global dictionary, while KCRC operates on the reduced kernel matrix without referring to original features.

The high level intuition of LCD is that local dictionary atoms are typically the most important and informative samples. Looking into these representative exemplars often brings even more gains than globally considering all samples together. It is not hard to see similar concepts and link the connections. For example, if the query is located near decision boundary, then these local atoms play the role similar to support vectors in an SVM, or in an extreme case, exemplars in an exemplar SVM [17]. In a model recommendation system, selecting the most responsive models instead of all models has been reported to give gains [15]. In fact, SRC also seeks to use only few exemplar atoms in the dictionary. Yet the proposed method is able to run much faster with even better performance. In the extreme scenario, if  $K$  equals the number of atoms, then the proposed KCRC-LCD degenerates to regular KCRC with global dictionary. If  $K$  equals 1, KCRC-LCD degenerates to the simplest nearest neighbor classifier.

In this paper, we specifically focus on the application of our proposed framework to the image classification/visual categorization problem, demonstrating its robust performance with a comprehensive series of image classification tasks. Image classification is among the most fundamental computer vision problems where each image is labeled with a certain or multiple categories/tags. Though great advance has been achieved, much is pending to be done since the current state-of-the-art approaches are far from being able to achieve human-level per-

formance, particularly in handling cluttered, complicated scenarios and inferring abstract concepts. Such gap between the machine and human remains an open challenge, motivating us to exploit more discriminative and efficient image classifiers.

The outline of the paper is as follows. Section II discusses related work of KCRC-LCD and presents our main contributions. In Section III, necessary preliminaries are briefly introduced. Section IV elaborates the formulation of KCRC-LCD and discusses some important details. The locality constrained dictionary is proposed and discussed in Section V. Experimental results are provided and discussed in Section VI, followed by concluding remarks in Section VII.

## 2. Related Work

Pioneering work for combining kernel technique to SRC is proposed in [18, 14, 19]. Gao et al. first proposed the idea of kernel-based SRC with a promising experimental results. Zhang et al. [14] further unified the mathematical model [18] to a generic kernel framework and conducted more comprehensive experiments to evaluate the performance. To overcome the shortcoming of handling data with the same direction distribution, Liu et al. [15] addressed the problem of kernel collaborative representation. The authors presented a smooth formulation to incorporate kernel function into the CRC model. A practical application of kernel CRC in vehicle logo recognition was further discussed in [20]. We happened to notice that a very recent work [21] proposed a similar idea by combining the column-generation kernel to CRC for hyperspectral image classification. It should be pointed out, however, that both the formulations as well as the applications are significantly different when dealing with the high dimension in kernel space. In terms of application, [21] formulated the kernel collaborative representation on pixel-level tasks for hyperspectral images while our method focus on image-level classification. Significant differences also exist in the formulation: [21] incorporated the kernel function with column generation without considering dimensionality reduction in kernel space (possibly due to the characteristics of the hyperspectral classification task). On the contrary, our method combines the CRC with kernel function in a strategy similar to KPCA and kernel Fisher discriminant analysis (KFDA). Moreover, a series of dimensionality reduction approaches have been taken into account in our generalized formulation. In general, we aim at extending the idea of KCRC by further improving formulation details of KCRC and presenting specific methods to perform dimensionality reduction in kernel space.

We also noticed that [22] presented a similar idea of constructing locally adaptive dictionary, but such dictionary pruning strategy was only applied in the standard CRC framework instead of the kernel CRC framework. As one shall see, the proposed KCRC-LCD is not a trivial extension of CRC-LCD by combining KCRC with locality constrained dictionary, but a well-motivated and appealing framework in which LCD and KCRC are theoretically linked by kernelizing the distance used in LCD. In addition, the kernelization in conjunction with LCD

not only brings scalability in terms of data set size, but also further extends its scalability to feature dimensionality. We will show that the locally adaptive dictionary in [22] is a special case of our proposed LCD.

### 3. Preliminaries

#### 3.1. Collaborative Representation Classification

The principle of CRC [11] is briefly presented in this section. In CRC, the dictionary  $\mathbf{D}$  is constructed by all training samples and a test sample  $\mathbf{y}$  is coded collaboratively over the whole dictionary, which is the essence of CRC.

Let  $\mathbf{D}$  be the dictionary that is a set of  $k$ -class training samples (the  $i$ th class with  $n_i$  samples), i.e.,  $\mathbf{D} = \{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_k\} \in \mathbb{R}^{m \times n}$  where  $n = \sum_{j=1}^k n_j$  and  $m$  is the feature dimension. Here, the dictionary associated with the  $i$ th class is denoted by  $\mathbf{D}_i = \{\mathbf{d}_1^{[i]}, \mathbf{d}_2^{[i]}, \dots, \mathbf{d}_{n_i}^{[i]}\} \in \mathbb{R}^{m \times n_i}$  in which  $\mathbf{d}_j^{[i]}$  - also called atom - stands for the  $j$ th training image in the  $i$ th class. Representing the query sample  $\mathbf{y}$  can be accomplished by solving  $\mathbf{x}$  in the following optimization model:

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \min_{\mathbf{x}} \|\mathbf{x}\|_{l_p} \\ \text{subj. to } \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_{l_q} &\leq \varepsilon \end{aligned} \quad (1)$$

where  $\varepsilon$  is a small error constant. After Lagrangian formulation, the model of CRC can be formulated as

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} (\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_{l_q} + \mu \|\mathbf{x}\|_{l_p}) \quad (2)$$

where  $\mu$  is the regularization parameter and  $p, q \in \{1, 2\}$ . The combinations of  $p, q$  lead to different instantiations of CRC model. For instance, SRC method is under the condition of  $p=1, q \in \{1, 2\}$  and different settings of  $q$  are used to handle classification with or without occlusion. Similar to SRC, CRC determines the class label via reconstruction residuals:

$$\text{identity}(\mathbf{y}) = \arg \min_i (\|\mathbf{y} - \mathbf{D}_i \hat{\mathbf{x}}_i\|_2 / \|\hat{\mathbf{x}}_i\|_2). \quad (3)$$

In fact, the key to reduce the computational complexity is to reasonably set the value of  $p, q$ . Based on different combinations of  $p, q$ , two CRC algorithms were proposed in [11, 23]. One is the CRC regularized least square (CRC-RLS) algorithm with  $p=2, q=2$ . The other is the robust CRC (RCRC) algorithm with  $p=2, q=1$ . [11] concluded that sparsity of signals is useful but not crucial for face recognition and proved that the collaborative representation mechanism does play an important role.

#### 3.2. Kernel Technique

In machine learning, kernel methods refer to a class of algorithms for pattern analysis, whose best known members are the SVM [24, 13], KPCA [12] and KFDA [25]. The general task of pattern analysis is to find and study general types of relations (for example clusters, rankings, principal components, correlations, classifications) in datasets. For many algorithms that solve these tasks, the data in raw representation have to be explicitly transformed into feature vector representations via a

user-specified feature map: in contrast, kernel methods require only a user-specified kernel, i.e., a similarity function over pairs of data points in raw representation. Via kernels, we can easily generalize a linear classifier to a nonlinear one, generating a reasonable decision boundary and consequently enhancing the discrimination power.

Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional, implicit feature space without computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors. The amazing part of kernel function is that it surpasses the direct calculation in the feature space and performs the classification in the reproducing kernel Hilbert space (RKHS), boosting the classification performance.

Algorithms that are capable of operating with kernels include the kernel perceptron [26, 27, 28], SVM [24, 13], Gaussian processes [29], principal components analysis (PCA) [12], canonical correlation analysis [30], spectral clustering [31] and many others. In general, any linear model can be transformed into a non-linear model by applying the kernel trick: replacing its features by a kernel function.

### 4. Proposed KCRC Approach

#### 4.1. Formulation of KCRC

To overcome the shortcoming of CRC in handling data with the same direction distribution, kernel technique is smoothly combined with CRC. Kernel function is used to create a non-linear mapping mechanism  $\mathbf{v} \in \mathbb{R} \mapsto \phi(\mathbf{v}) \in \mathbb{H}$  in which  $\mathbb{H}$  is a unique associated RKHS. If every sample is mapped into higher dimensional space via transformation  $\phi$ , the kernel function is written as

$$K(\mathbf{v}', \mathbf{v}'') = \langle \phi(\mathbf{v}'), \phi(\mathbf{v}'') \rangle = \phi(\mathbf{v}')^T \phi(\mathbf{v}'') \quad (4)$$

where  $\mathbf{v}', \mathbf{v}''$  are different samples and  $\phi$  denotes the implicit nonlinear mapping associated with the kernel function  $K(\mathbf{v}', \mathbf{v}'')$ . There are some empirical kernel functions satisfying the Mercer condition such as the linear kernel  $K(\mathbf{v}', \mathbf{v}'') = \mathbf{v}'^T \mathbf{v}''$  and radial basis function (RBF) kernel  $K(\mathbf{v}', \mathbf{v}'') = \exp(-\beta \|\mathbf{v}' - \mathbf{v}''\|_2^2)$ . According to [32], the distance function for similarity measurement, designed to construct the LCD, can be transformed in a straightforward way to the kernel for KCRC, via the linear kernel function:

$$\begin{aligned} K(\mathbf{v}', \mathbf{v}'') &= \langle \phi(\mathbf{v}'), \phi(\mathbf{v}'') \rangle = \langle \mathbf{v}', \mathbf{v}'' \rangle \\ &= \frac{1}{2} (\langle \mathbf{v}', \mathbf{v}' \rangle + \langle \mathbf{v}'', \mathbf{v}'' \rangle - \langle \mathbf{v}' - \mathbf{v}'', \mathbf{v}' - \mathbf{v}'' \rangle) \\ &= \frac{1}{2} (\text{Dist}(\mathbf{v}', 0) + \text{Dist}(\mathbf{v}'', 0) - \text{Dist}(\mathbf{v}', \mathbf{v}'')) \end{aligned} \quad (5)$$

where  $\text{Dist}$  is the carefully designed distance function, and the location of the origin(0) does not affect the result [32]. Vari-

ous ways of transforming a distance function into a kernel are possible [33], i.e.,  $K(\mathbf{v}', \mathbf{v}'')$  can be  $\exp(-\text{Dist}(\mathbf{v}', \mathbf{v}'')/\beta^2)$ .

It is learned in [13] that the sample feature nonlinearly transformed to high dimensional space becomes more separable. Most importantly, the same direction distribution of data can be avoided in kernel space. However, mapping to high dimensional space makes CRC model harder to solve, so we need to perform dimensionality reduction in the kernel feature space. The nonlinear mapping mechanism is

$$\mathbf{y} \in \mathbb{R}^m \mapsto \phi(\mathbf{y}) = [\phi_1(\mathbf{y}), \phi_2(\mathbf{y}), \dots, \phi_s(\mathbf{y})] \in \mathbb{R}^s \quad (6)$$

where  $\phi(\mathbf{y})$  is the high dimensional feature corresponding to the sample  $\mathbf{y}$  and  $s \gg m$ . Then we define a universal label  $[k]$  for  $\mathbf{d}_j^{[i]}$  that denotes its position in the global dictionary, satisfying  $k = j + \sum_{l=1}^{i-1} n_l$ . For conciseness, we only preserve the universal label, representing atom as  $\mathbf{d}_{[k]}$ . According to the nonlinear mapping mechanism, the original dictionary  $\mathbf{D}$  has a much higher dimension:  $\Phi = \{\phi(\mathbf{d}_{[1]}), \phi(\mathbf{d}_{[2]}), \dots, \phi(\mathbf{d}_{[n]})\} \in \mathbb{R}^{s \times n}$ , and the test sample becomes  $\phi(\mathbf{y}) = \Phi \mathbf{x}$ . The KCRC model is formulated as

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \min_{\mathbf{x}} \|\mathbf{x}\|_{l_p} \\ \text{subj. to } \|\phi(\mathbf{y}) - \Phi \mathbf{x}\|_{l_q} &\leq \varepsilon. \end{aligned} \quad (7)$$

However, Eq. (7) is harder to solve than Eq. (1) because the high dimensionality results in high complexity. A dimensionality reduction matrix  $\mathbf{R}$ , namely a projection matrix, can be constructed by utilizing the methodology in KPCA [12] and KFDA [25]. With the matrix  $\mathbf{R} \in \mathbb{R}^{s \times c}$ , we derive

$$\mathbf{R}^T \phi(\mathbf{y}) = \mathbf{R}^T \Phi \mathbf{x} \quad (8)$$

where  $\mathbf{R}$  is related to kernelized samples. In both KPCA and KFDA, each column vector in  $\mathbf{R}$  is a linear combination of kernelized samples, which is also adopted in KCRC. Namely

$$\mathbf{R} = \Phi \Psi = \{\phi(\mathbf{d}_{[1]}), \dots, \phi(\mathbf{d}_{[n]})\} \cdot \{\psi_1, \dots, \psi_c\} \quad (9)$$

where  $\mathbf{R} = \{\mathbf{R}_1, \dots, \mathbf{R}_s\}$  and  $\psi_i$  is the  $n$ -dimensional linear projection coefficients vector:  $\mathbf{R}_i = \sum_{j=1}^n \psi_{i,j} \phi(\mathbf{d}_{[j]}) = \Phi \psi_i$ . Moreover,  $\Psi \in \mathbb{R}^{n \times c}$  is also called pseudo-transformation matrix [14]. Then we put Eq. (9) into Eq. (8) and obtain

$$(\Phi \Psi)^T \phi(\mathbf{y}) = (\Phi \Psi)^T \Phi \mathbf{x} \quad (10)$$

from which we get  $\Psi^T \mathbf{K}(\mathbf{D}, \mathbf{y}) = \Psi^T \mathbf{G} \mathbf{x}$ , where  $\mathbf{K}(\mathbf{D}, \mathbf{y}) = [K(\mathbf{d}_{[1]}, \mathbf{y}), \dots, K(\mathbf{d}_{[n]}, \mathbf{y})]^T$ .  $\mathbf{G} (G_{ij} = K(\mathbf{d}_{[i]}, \mathbf{d}_{[j]}))$ , also equal to  $\Phi^T \Phi$ , is defined as the kernel Gram matrix that is symmetric and positive semi-definite according to Mercer's theorem. Since  $\mathbf{G}$  and  $\mathbf{K}(\mathbf{D}, \mathbf{y})$  are given a priori, dimensionality reduction requires to find  $\Psi$  instead of  $\mathbf{R}$ . Several methods were introduced in [14, 25, 12] to determine the pseudo transformation matrix  $\Psi$ . We will also further introduce the selection of matrix  $\Psi$  in the next subsection. Note that, if  $\Psi$  is an identity matrix, no dimensionality reduction is applied. Particularly,  $\Psi$  can also be a random projection matrix to achieve dimensional-

ity reduction. After substituting the equivalent kernel function constraint, we can derive

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \min_{\mathbf{x}} \|\mathbf{x}\|_{l_p} \\ \text{subj. to } \|\Psi^T \mathbf{K}(\mathbf{D}, \mathbf{y}) - \Psi^T \mathbf{G} \mathbf{x}\|_{l_q} &< \varepsilon \end{aligned} \quad (11)$$

which is the model of KCRC approach. Additionally, a small perturbation would be added to  $\Psi^T \mathbf{G}$  if the norm of a column is close to 0. Another form of KCRC model is expressed as

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} (\|\Psi^T \mathbf{K}(\mathbf{D}, \mathbf{y}) - \Psi^T \mathbf{G} \mathbf{x}\|_{l_q} + \mu \|\mathbf{x}\|_{l_p}) \quad (12)$$

from which we could derive two specific algorithms. With  $p=2, q=2$ ,  $\mathbf{x}$  can be solved at the cost of low computational complexity. The regularized least square algorithm is used to solve the optimization problem (Algorithm 1). Handling images with occlusion and corruption, we can set  $p=2, q=1$  for robustness, making the first term a  $l_1$  regularized one. Let  $\mathbf{e} = \Psi^T \mathbf{K}(\mathbf{D}, \mathbf{y}) - \Psi^T \mathbf{G} \mathbf{x}$  and  $p=2, q=1$ . Eq. (11) is rewritten as

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \min_{\mathbf{x}} (\|\mathbf{e}\|_1 + \mu \|\mathbf{x}\|_2) \\ \text{subj. to } \Psi^T \mathbf{K}(\mathbf{D}, \mathbf{y}) &= \Psi^T \mathbf{G} \mathbf{x} + \mathbf{e} \end{aligned} \quad (13)$$

which is a constrained convex optimization problem that can be solved by the augmented Lagrange multiplier (ALM) method [34, 35] as shown in Algorithm 2.

#### 4.2. Determining the Pseudo-transformation Matrix for Dimensionality Reduction

This subsection reviews several typical methods that are proposed in [14] to determine the pseudo-transformation matrix  $\Psi$  for dimensionality reduction. Moreover, we also present a graph preserving method that has not been utilized to construct pseudo-transformation matrix in previous work.

##### 4.2.1. KPCA

Following the methodology in KPCA, the pseudo-transformation vectors  $\psi_i \in \mathbb{R}^n$  refer to normalized eigenvectors corresponding to nonzero eigenvalues (or greater than a threshold) which can be obtained from the following eigenvalue problem [12]:

$$n \lambda \psi_i = \mathbf{G} \psi_i \quad (14)$$

where  $\psi_i$  is normalized to satisfy  $\lambda_i \psi_i^T \psi_i = 1$ . Eq. 14 can be easily solved by singular value decomposition (SVD) method.  $\Psi$  is equal to  $\{\psi_1, \psi_2, \dots, \psi_c\}$ .

##### 4.2.2. KFDA

For KFDA [25],  $\Psi \in \mathbb{R}^{n \times c}$  is the solution of the optimization problem shown as follows:

$$\arg \max_{\Psi} \frac{\text{tr}(\Psi \mathbf{S}_b^G \Psi)}{\text{tr}(\Psi \mathbf{S}_w^G \Psi)} \quad (15)$$

where  $\text{tr}(\cdot)$  denotes the trace of a matrix, and  $\mathbf{S}_b^G, \mathbf{S}_w^G$  stand for quasi within-class and between-class scatter matrices respectively.



#### 4.2.3. Random Projection

[14] also proposed a simple and practical random dimensionality reduction method. Since random projection can not be performed in the RHKS, we can make  $\Psi$  a Gaussian random matrix to reduce the dimensionality. Random projection can be viewed as a less-structured counterpart to classic dimensionality reduction methods like PCA and FDA. In other word, the critical information will be preserved in a less-structured way.

#### 4.2.4. Identity

Particularly, the pseudo-transformation matrix  $\Psi$  can be defined as an identical matrix with ones on the main diagonal and zeros elsewhere, which indicates no dimensionality reduction is performed in the RHKS. The method is the most simple way for dimensionality reduction in KCRC, but it is effective at most time, especially in KCRC-LCD. LCD is usually constructed in relatively small size compared to the training sets, so we do not always need to perform dimensionality reduction in kernel space whose dimension is equal to the dictionary size. Thus, in the classification experiments on public database, we simply use identity matrix as  $\Psi$ .

#### 4.2.5. Graph

Further, we propose a graph preserving dimensionality reduction method for the pseudo-transformation matrix  $\Psi$ . In the light of [36], we first construct a weighted graph with  $n$  nodes ( $n$  is the dictionary size, one node represents one atom in the dictionary). Then we put an edge between node  $i$  and  $j$  if they are close enough. Usually, there are two methods to find the nodes that we use to construct the graph. The first is  $\epsilon$ -neighborhoods in which node  $i$  and node  $j$  are connected by an edge if  $\|d_i - d_j\|^2 < \epsilon$ . The second is  $n$ -nearest neighbors in which node  $i$  and  $j$  connected by an edge if  $d_i$  is among  $n$  nearest neighbors of  $d_j$  or  $d_j$  is among  $n$  nearest neighbors of  $d_i$ . After constructing the weighted graph which contains the similarity information among atoms, we choose a measure for the weight. In [36], the following weight measure between two connected nodes is formulated as

$$W_{ij} = \exp\left(\frac{\|d_i - d_j\|^2}{t}\right) \quad (16)$$

besides which, there is another simple weighting method that  $W_{ij} = 1$  if and only if vertices  $i$  and  $j$  are connected by an edge. In order to group the connected nodes and separate the distant nodes as much as possible, the object function is defined as

$$\sum_{ij} (g_i - g_j)^2 W_{ij} = 2g^T L g \quad (17)$$

where  $g_i$  ( $0 \leq i \leq n$ ) is the map from the graph to the real sample and  $L$  is the Laplacian matrix satisfying  $L = D - W$  in which  $D_{ii} = \sum_j W_{ij}$ . Laplacian matrix is a symmetric, positive semi-definite matrix which can be thought of as an operator on functions defined on vertices of the graph. Then we

can formulate the minimization problem as

$$\begin{aligned} & \arg \min_g g^T L g \\ & \text{subj. to } g^T D g = 1 \end{aligned} \quad (18)$$

which is equivalent to the solution of the following generalized eigenvalue decomposition problem:

$$Lg = \lambda Dg \quad (19)$$

which is similar to the optimization problem in PCA. We let  $g_0, g_1, \dots, g_{n-1}$  be the solutions of Eq.19, sorted according to their eigenvalues with  $g_0$  having the smallest eigenvalue (actually it is zero). After performing normalization on  $\{g_1, g_2, \dots, g_c\}$ , we take them as the pseudo-transformation matrix  $\Psi$ , namely

$$\Psi = \{g_1, g_2, \dots, g_c\}. \quad (20)$$

The motivation of this dimensionality approach is quite intuitive. We construct  $\Psi$  with the graph constraint in order to combine the graph information, or to be more accurate, the similarity relation among atoms into the kernel space (after dimensionality reduction).

#### 4.2.6. Further Discussion

We present four methods to perform the dimensionality reduction in the kernel space. Reducing the dimensionality in the kernel space brings several gains such as lowering the computational cost and enhancing the discrimination power. There also exist a number of other ways to perform the dimensionality reduction in the kernel space, namely construct the matrix  $\Psi$ . Empirically, if the rank of matrix  $\Psi$  stays unchanged, then different construction of  $\Psi$  will not lead to dramatical difference in classification accuracy. Thus, the matrix  $\Psi$  is not very crucial to the classifier, which is supported by the experiments conducted in [14]. Instead, the rank of the matrix  $\Psi$  plays a crucial part in classification accuracy. This is why even using random matrix as  $\Psi$  still serves our classifier well. In Section IV, we conduct relevant experiments to study what the selection of the matrix  $\Psi$  will do to the classification accuracy.

#### 4.3. Practical KCRC Algorithms

There are two algorithms designed for KCRC. For normal situations,  $p, q$  are both set as 2. The regularized least square algorithm is adopted to solve the model with  $p, q=2$ . Specifically, we derive the new dictionary  $D' = \Psi^T G$  and define  $P'$  as the coding basis in kernel CRC-RLS (KCRC-RLS). Namely

$$P' = ((\Psi^T G)^T (\Psi^T G) + \mu \cdot I)^{-1} (\Psi^T G)^T \quad (21)$$

where  $\mu$  is a small constant. The query sample is transformed to  $\Psi^T K(D, y)$ . Apparently,  $P'$  is independent of  $y'$  so it can be pre-calculated. When a query  $y$  comes, the query is first transformed to the kernel space via  $y' = \Psi^T K(D, y)$  and then can be simply projected onto the coding basis  $P'$  via  $P'y'$ . In the decision making stage, class-specified representation residual  $\|y' - D'_i \hat{x}_i\|_2$  is used for classification. Further, a  $l_2$  norm

term  $\|\hat{x}_i\|_2$  is added for more discriminative classification. The specific algorithm of KCRC-RLS is shown in Algorithm 1.

For high level corruption and occlusion, kernel robust CRC (KRCRC) algorithm ( $p=2, q=1$ ) can be applied. Note that,  $D''=\Psi^T G$  and  $P_k''$  are designed as the new dictionary and coding basis in kernel space respectively.

$$P'' = ((\Psi^T G)^T (\Psi^T G) + 2\mu/\sigma_k \cdot I)^{-1} (\Psi^T G)^T \quad (22)$$

where  $\mu, \sigma_k$  are small constants. The augmented Lagrangian function used for the optimization in Eq. (13) is formulated as

$$L_\sigma(e, x, z) = \|e\|_1 + \mu \|x\|_2^2 + \langle z, y'' - D''x - e \rangle + \frac{\sigma}{2} \|y'' - D''x\|_2^2 \quad (23)$$

where  $\sigma$  is a positive constant that is the penalty for large representation error, and  $z$  is a vector of Lagrange multiplier. The ALM method iteratively estimates  $e, x$  for the Lagrange multiplier  $z$  via the following minimization:

$$(e_{k+1}, x_{k+1}) = \arg \min_{e, x} L_{\sigma_k}(e, x, z_k) \quad (24)$$

where  $z_{k+1} = z_k + \sigma_k(y'' - D''x - e)$ . According to [23, 34], this iteration will converge to an optimal solution for Eq. (13) if  $\{\sigma_k\}$  is a monotonically increasing sequence.

The minimization process in Eq. (24) can be implemented by optimizing  $e, x$  alternatively and iteratively:

$$\begin{aligned} x_{k+1} &= \arg \min_x L_{\sigma_k}(x, e_k, z_k), \\ e_{k+1} &= \arg \min_e L_{\sigma_k}(x_{k+1}, e, z_k), \end{aligned} \quad (25)$$

which has the closed-form solution as follows:

$$\begin{aligned} x_{k+1} &= (D''^T D'' + 2\mu/\sigma_k I)^{-1} D''^T (y'' - e_k + z_k/\sigma_k) \\ &= P_k''(y'' - e_k + z_k/\sigma_k), \\ e_{k+1} &= S_{1/\sigma_k}(y'' - D''x_{k+1} + z_k/\sigma_k), \end{aligned} \quad (26)$$

where the function  $S_\alpha, \alpha \geq 0$  is the soft-thresholding (shrinkage) operator given by

$$S_\alpha(h) = \begin{cases} h - \alpha, & \text{if } h \geq \alpha \\ h + \alpha, & \text{if } h \leq -\alpha \\ 0, & \text{otherwise} \end{cases} \quad (27)$$

If  $h$  represents a  $n$ -dimensional vector, then  $S_\alpha(h)$  is given by  $\{S_\alpha(h_1), S_\alpha(h_2), \dots, S_\alpha(h_n)\}$ . Similar to the KCRC-RLS, the coding basis  $P_k''$  is independent of  $y''$  for the given  $\sigma_k$ , so the set of projection matrices  $\{P_k\}$  can also be pre-calculated. Once a query sample  $y$  comes, it is first transformed in the kernel space via  $\Psi^T K(D, y)$  and then projected onto  $P_k''$  via  $P_k'' y''$ . After performing the iterative minimization above, a classification strategy similar to KCRC-RLS is applied in KR-CRC. Details of KRCRC is given in Algorithm 2.

---

#### Algorithm 1: KCRC-RLS

---

1. Normalize the columns of  $D' = \Psi^T G$  to unit  $l_2$ -norm.
  2. Represent  $y' = \Psi^T K(D, y)$  over dictionary  $D'$  by  $\hat{x} = P' y'$  where  $P' = (D'^T D' + \mu I)^{-1} D'^T$ .
  3. Obtain the regularized residuals  $r_i = \|y' - D'_i \hat{x}_i\|_2 / \|\hat{x}_i\|_2$  where  $\hat{x}_i$  is the coding coefficients associated with class  $i$  over  $P'$ .
  4. Output the identity of  $y'$  (class label) as  $identity(y') = \arg \min_i(r_i)$ .
- 

---

#### Algorithm 2: KRCRC

---

1. Normalize the columns of  $D'' = \Psi^T G$  to unit  $l_2$ -norm.
  2. Input  $y'' = \Psi^T K(D, y)$ ,  $x_0, e_0, k=1$  and  $\tau > 0$ .
  3. Proceed if  $\|x_{k+1} - x_k\|_2 > \tau$  is true. If not, output  $\hat{e}, \hat{x}$  and go to step 5.
  4. Do the following iteration:  $x_{k+1} = P_k''(y'' - e_k + z_k/\sigma_k)$   $e_{k+1} = S_{1/\sigma_k}(y'' - D''x_{k+1} + z_k/\sigma_k)$   $z_{k+1} = z_k + \sigma_k(y'' - D''x_{k+1} - e_{k+1})$  where  $P_k'' = (D''^T D'' + 2\mu/\sigma_k I)^{-1} D''^T$  and  $S_\alpha, \alpha \geq 0$  is the shrinkage coefficient.  $k \leftarrow k + 1$  and go to step 3.
  5. Represent  $y''$  over dictionary  $D''$  by the converged  $x$ .
  6. Obtain the regularized residuals  $r_i = \|y'' - D''_i \hat{x}_i\|_2 / \|\hat{x}_i\|_2$  where  $\hat{x}_i$  is the coding coefficients related to class  $i$ .
  7. Output the identity of  $y''$  (class label) as  $identity(y'') = \arg \min_i(r_i)$ .
- 

## 5. On the Locality Constrained Dictionary

This section elaborates the locality constrained dictionary. Additionally, we present some typical distances used in LCD for similarity measurement and further introduce a distance fusion model, followed by the introduction of the KCRC method combined with LCD, termed as KCRC-LCD.

### 5.1. Locality Constrained Dictionary

Most collaborative representation based methods [23, 37, 38] employ all the high-dimensional training samples as the global dictionary. They may work fine when the global dictionary is small, but the classification becomes intractable with increasingly more training samples. To tackle with this problem, we propose the LCD that utilizes the K-NN classifier to measure the similarities between the query sample and all atoms in the global dictionary, and then selects  $K$  nearest atoms as the local dictionary. The locality in LCD ensures discrimination, efficiency and robustness of KCRC. Compared to the locality constrained dictionary proposed in [39], we adopt a more straightforward way to constrain the locality, which needs no learning and training process, greatly reducing the computational cost in training. Under such locality constrained dictionary, scaling to a large number of categories dose not require

adding new features, because the discriminative distance function need only be defined for similar enough samples. From biological and psychological perspective, similarity between samples is the most important criteria to recognize and classify objects for human brains. So intuitively speaking, the proposed locality constrained dictionary with various optional discriminative distances makes our KCRC approach more scalable, discriminative, efficient, and most importantly, free from the curse of high-dimensional feature space. Moreover, the kernel idea within KCRC well suits the idea of locality both mathematically and experimentally.

Define  $Dist(\mathbf{d}', \mathbf{d}'')$  as the distance metric between atom  $\mathbf{d}'$  and  $\mathbf{d}''$ . To be simple, we adopt the  $l_2$  distance as example in formulation, namely  $Dist(\mathbf{d}', \mathbf{d}'') = \|\mathbf{d}' - \mathbf{d}''\|_2$ . We need to calculate the distance between every atom  $\mathbf{d}_{[k]}$  and the query sample  $\mathbf{y}$  first. Then the LCD can be obtained via the following optimization:

$$\begin{aligned} \arg \min_{\{t_1, t_2, \dots, t_K\}} \sum_{m=1}^K Dist(\mathbf{d}_{[t_m]}, \mathbf{y}) \\ \text{subj. to } 1 \leq t_i \neq t_j \leq n, \text{ for } \forall i \neq j \end{aligned} \quad (28)$$

where  $\mathbf{d}_{[t_1]}, \mathbf{d}_{[t_2]}, \dots, \mathbf{d}_{[t_K]}$  denote different atoms in the global dictionary. In fact, to solve Eq. (28) is to find the  $K$  atoms that are located nearest to the query sample. As a result, the LCD is obtained as  $\mathbf{D}_{lc} = \{\mathbf{d}_{[t_1]}, \mathbf{d}_{[t_2]}, \dots, \mathbf{d}_{[t_K]}\}$ . Moreover, the computational complexity of solving Eq. (28) is  $O(n \log n)$ , which is efficient enough to perform in large-scale image databases. Note that, when  $K=n$ , KCRC-GD becomes a special case of KCRC-LCD.

## 5.2. Discriminative Distances for Similarity Measure

In the previous subsection, we simply use the  $l_2$  distance as an example. In fact, there are many discriminative distances for similarity measurement. Several well-performing distances are introduced in [33], i.e., Mahalanobis distance,  $\chi^2$  distance [40], marginal distance [41], tangent distance [42], shape context based distance [43] and geometric blur based distance [44], etc. Each can be used to measure the similarity in order to construct a well-performing LCD. These distances can either be used alone or used in conjunction with each other, making the LCD flexible and adaptive. We will review some of the discriminative distances in this subsection.

### 5.2.1. General Pixel Similarity Measure

We consider several classical pixel distance metrics below. Euclidean distance ( $l_2$  distance) is the most popular similarity measure. It is simple yet effective in certain situations and defined as

$$Dist(\mathbf{d}', \mathbf{d}'') = \|\mathbf{d}' - \mathbf{d}''\|_2. \quad (29)$$

City block distance, also known as Manhattan distance, assumes that it is only possible to travel along pixel grid lines from one pixel to another. This distance metric is defined as

$$Dist(\mathbf{d}', \mathbf{d}'') = \|\mathbf{d}' - \mathbf{d}''\|_1. \quad (30)$$

Chessboard distance metric assumes that you can make moves on the pixel grid as if you were a King making moves in chess, i.e. a diagonal move counts the same as a horizontal move. The metric is given by:

$$Dist(\mathbf{d}', \mathbf{d}'') = \|\mathbf{d}' - \mathbf{d}''\|_\infty. \quad (31)$$

There are a lot of other general pixel distances that can be utilized in our framework, such as correlation distance, Mahalanobis distance, etc.

### 5.2.2. Texture Similarity Measure

We present some texture similarity measure as examples below. The  $\chi^2$  distance is proposed in [40] for texture similarity measure. The main idea of the  $\chi^2$  distance is to construct a vocabulary of 3D textons by clustering a set of samples. Associated with each texton is an appearance vector which characterizes the local irradiance distribution. The similarity can be measured by characterizing samples with these 3D textons.

In the view of statistics, marginal distance [41] is another version of the  $\chi^2$  distance. They both measure the difference between two joint distribution of texture response. The difference is that marginal distance metric simply sums up the distance between response histograms from each filter while the  $\chi^2$  distance metric measures the similarity of the two joint distribution by comparing the histogram of textons.

As another texture similarity measure initially used for handwritten digits recognition, Tangent distance [42] is defined to compute the minimum distance between the linear surfaces that best approximate the non-linear manifolds of different sample categories. These linear surfaces, which are crucial to Tangent distance, are derived from the images by including perturbations from small affine transformation of the spatial domain and change in the thickness of pen-stroke.

### 5.2.3. Shape Similarity Measure

The shape in an image can be represented by a set of points, with a descriptor at a fixed point to measure the relative position to that point. These descriptors are iteratively matched using a deformation model. Shape context based distance [43] is derived from the discrepancy left in the final matched shapes and a score that denotes how far the deformation is from an affine transformation. Various shape descriptors can be defined on a gray scale image, for example, the shape context descriptor on the edge map [45], the SIFT descriptor [46], and the geometric blur descriptor [44], optimized local shape descriptor [47], etc.

### 5.2.4. Unified Similarity Measure

We use a simple and intuitive method to combine general pixel similarity, texture similarity and shape similarity into a unified locality constrained dictionary. Assume that we need to construct a LCD with size  $K$  out of total  $N$  training samples. First, we enforce locality to dictionary via general pixel similarity, texture similarity and shape similarity, obtaining three LCD with size  $K$ :  $\mathbf{D}_{lc}^{[pixel]}$ ,  $\mathbf{D}_{lc}^{[texture]}$  and  $\mathbf{D}_{lc}^{[shape]}$ . From the sets perspective, these three dictionaries constructed via different similarity measure can be viewed in Venn diagram as shown

in Fig. 1. Specifically, an atom in dictionary represents an element in a set, so a LCD can be regarded as a set with  $K$  elements. According to the demand of the given task, we can use different combinations of similarity measures to construct the LCD. Mathematically, we achieve the combination of similarity measures by getting the union of the corresponding LCDs, i.e.,  $D_{lc} = D_{lc}^{[texture]} \cup D_{lc}^{[shape]}$  for the combination of texture and shape similarity measure. With unified similarity measure, the distance metric used in the kernel function becomes a linear combination of the distances that are utilized to construct the new LCD. Normally, we suppose the weight of each distance metric in the unified distance is equal. However, to use unified similarity measure could add to computational cost, so we do not recommend to use it under normal circumstance. Note that, the same type similarity measures can also be unified by the proposed method with similar procedure.

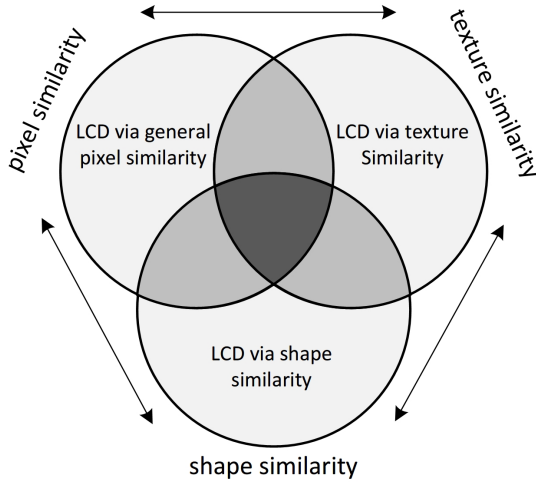


Figure 1. Venn diagram of LCDs constructed via general pixel similarity, texture similarity and shape similarity.

### 5.3. KCRC-LCD Algorithm

Intuitively, we use the locality constrained dictionary  $D_{lc}$  in place of global dictionary  $D$  and then perform the KCRC algorithm on  $D_{lc}$ . The KCRC-LCD algorithm is as follows:

---

#### Algorithm 3: Naive KCRC-LCD

---

1. Compute the distances between the query sample and all training samples, and pick the nearest  $K$  neighbors.
  2. If the  $K$  neighbors have the same labels, the query is labeled and exit; Else, construct the LCD  $D_{lc}$  with the  $K$  labeled neighbors and goto Step 3.
  3. Convert the pairwise distance into a kernel matrix via kernel trick and utilize KCRC approach with dictionary  $D_{lc}$  instead of the global dictionary  $D$ .
  4. Output the label of the query sample.
- 

The naive version of KCRC-LCD performs slowly because it has to compute the distances of the query to all training samples. Inspired by [33], we consider to boost the efficiency in the coarse-to-fine framework which is similar to the human perception procedure that human first perform a fast coarse pruning

and then recognize the object by details. The practical version of KCRC-LCD is as follows:

---

#### Algorithm 4: Practical KCRC-LCD

---

1. Compute the coarse distances (i.e. Euclidean distance) between the query sample and all training samples, and pick the nearest  $K_c$  neighbors. ( $K_c \geq K_f$ )
  2. Compute the fine distances between the query sample and pick the nearest  $K_f$  neighbors.
  2. If the  $K_f$  neighbors have the same labels, the query is labeled and exit; Else, construct the LCD  $D_{lc}$  with the  $K_f$  labeled neighbors and goto Step 3.
  3. Convert the pairwise distance into a kernel matrix via kernel trick and utilize KCRC approach with dictionary  $D_{lc}$  instead of the global dictionary  $D$ .
  4. Output the label of the query sample.
- 

## 6. Experiments and Results

### 6.1. Evaluation of Dimensionality Reduction in Kernel Space

We evaluate the dimensionality reduction from two aspects. First we compare the representation coefficients and reconstruction residuals that are obtained by SRC, CRC, KCRC with no dimensionality reduction (KCRC-Identity), KCRC with KPCA dimensionality reduction (KCRC-KPCA), KCRC with random projection (KCRC-RP) and KCRC with graph projection (KCRC-Graph). Second, we compare the recognition accuracy of these methods in extended Yale B face database.

We randomly select 38 images per person (32 person, totally 1216 images) in extended Yale B database [48] as training samples. For the computation of representation coefficients and reconstruction residuals, we use a single fixed test sample for better comparison. To simplify the experiment, we only use the global dictionary since the experiment focuses on the dimensionality reduction methods in kernel space. In the recognition accuracy test, we follow the same experiments settings as [3] by randomly selecting 38 images per person (32 person, totally 1216 images) and using the remaining images as test samples. The results are averaged over 20 times experiments.

Fig. 2 gives the representation coefficients and reconstruction residuals of SRC, CRC, KCRC-Identity, KCRC-KPCA, KCRC-RP and KCRC-Graph. We can see all of these five approaches tell the correct label (the first class) of the test sample. It can be obtain from Fig. 2 that KCRC-Identity and KCRC-Graph achieve better sparsity, similar to the representation coefficients of SRC. Moreover, the reconstruction residuals of all these approaches indicate the first class has the fewest reconstruction residual. Table. 1 shows the recognition accuracy of SRC, CRC, KCRC-Identity, KCRC-KPCA, KCRC-RP and KCRC-Graph on extended Yale B database. We can see CRC has the best recognition accuracy of these five methods. However, all these approaches are of the same level discrimination ability since the difference between the highest recognition rate and the lowest is less than 1%.



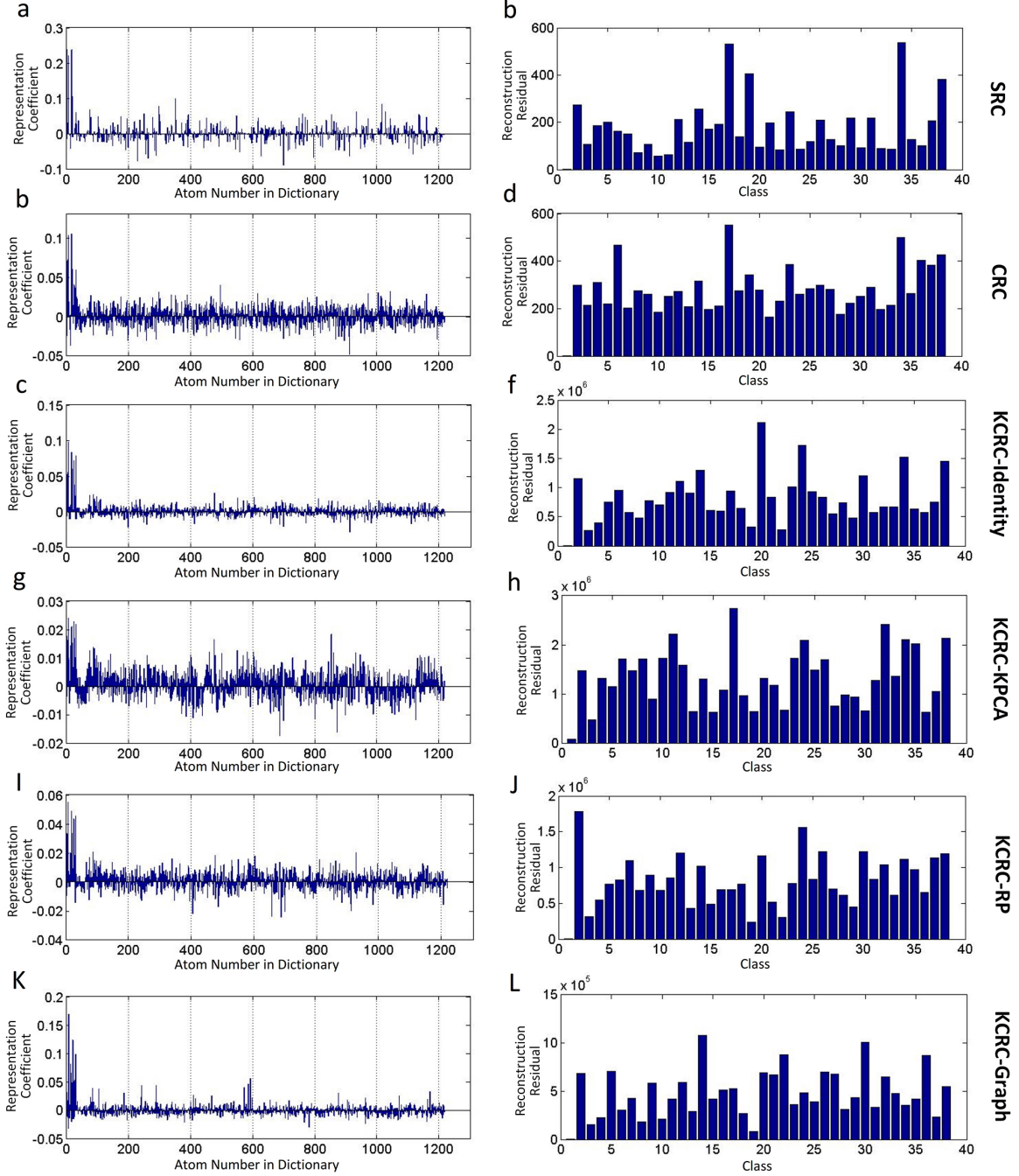


Figure 2. (a) Representation coefficients obtained by SRC. (b) Reconstruction Residuals obtained by SRC. (c) Representation coefficients obtained by CRC. (d) Reconstruction Residuals obtained by CRC. (e) Representation coefficients obtained by KCRC-Identity. (f) Reconstruction Residuals obtained by KCRC-Identity. (g) Representation coefficients obtained by KCRC-KPCA. (h) Reconstruction Residuals obtained by KCRC-KPCA. (i) Representation coefficients obtained by KCRC-RP. (j) Reconstruction Residuals obtained by KCRC-RP. (k) Representation coefficients obtained by KCRC-Graph. (l) Reconstruction Residuals obtained by KCRC-Graph.

## 6.2. Experiments on Data with the Same Direction Distribution

We evaluate the performance on data with the same direction distribution. In Fig. 3, we compare 3 classifiers: CRC-GD, KCRC-GD and KCRC-LCD. Two-class training data  $\mathbf{Q}$ ,  $\mathbf{W}$  with  $m$ -dimension are generated for the experiment. Each feature of  $\mathbf{Q}$ ,  $\mathbf{W}$  uniformly takes value from the interval  $[1, 3]$  and  $[-3, -1]$  respectively, corrupted by Gaussian noise with zero

mean and 0.15 variance. Then, we let  $m$  vary from 2 to 256 and perform the experiment. The results show that both KCRC-GD and KCRC-LCD can perfectly classify data with the same direction distribution while CRC performs poorly. This experiment shows both KCRC-GD and KCRC-LCD could handle data with special distribution, i.e. the same direction distribution in this case. Thus, kernel function makes our proposed ap-

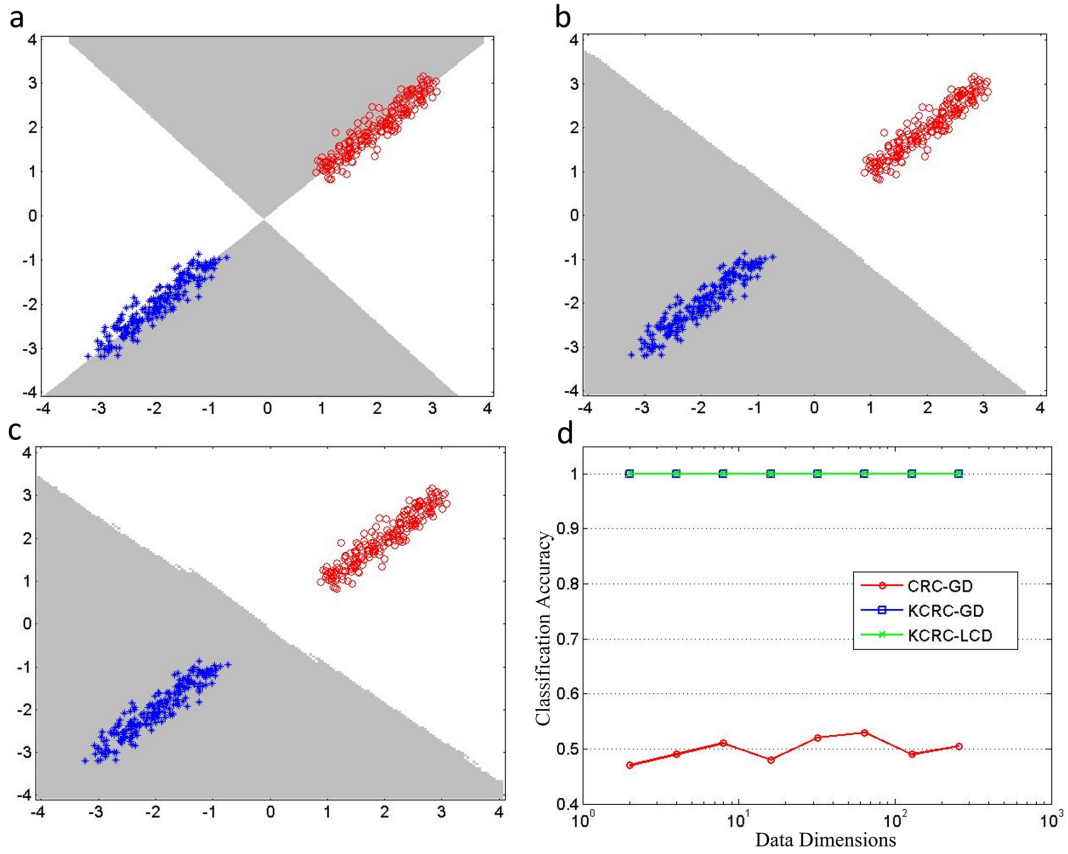


Figure 3. Performance comparison on data with the same direction distribution. Test samples are from the entire surface whose predicted labels are indicated by gray or white. 2-D decision boundaries obtained by (a) CRC with global dictionary (CRC-GD), (b) KCRC with global dictionary (KCRC-GD), (c) KCRC with locality constrained dictionary (KCRC-LCD). (d) Classification accuracy vs. dimensionality.

Table 1. Recognition results on extended Yale B database. 504 random projection features and the global dictionary (1216 atoms) are adopted. The results below are averaged over 10 times experiments.

Method	Accuracy(%)
SRC	97.15
CRC	97.67
KCRC-Identity	97.23
KCRC-KPCA	97.07
KCRC-RP	96.61
KCRC-Graph	97.35

proach more prepared for unknown data distribution than conventional CRC.

### 6.3. Experiments on Public Databases

This subsection evaluates our approach on public databases. Reliable results are obtained by 20 times repeated experiments with different random splits of the training and test images.

#### 6.3.1. MNIST

The MNIST database [49] of handwritten digits contains 60,000 samples (10 digits) for training and 10,000 for testing. For the experimental settings,  $l_2$  distance and  $28 \times 28$  raw pixel features are used for similarity measure and classification. We

evaluate our approach via different dictionary size 100, 200, 500, 700, 1000 and 1500, namely 10, 20, 50, 70, 100 and 150 samples for training per category. For settings of KCRC-LCD and CRC-LCD, we use the global dictionary of size 2000 (200 training samples per category) to generate the LCD and set  $K$  for LCD as 100, 200, 500, 700, 1000 and 1500 for comparison. Experimental results in Fig. 4(a) show KCRC-LCD has the best performance compared to CRC-LCD, CRC-GD and KCRC-GD in the MNIST database. From Fig. 4(b), it can be learned that  $K$  has slight impact on classification accuracy if the global dictionary is fixed (atom number of GD stays unchanged).

#### 6.3.2. Extended Yale B Faces

The extended Yale B database consists of 2414 frontal face images of 38 individuals [48]. The cropped  $192 \times 168$  face images are taken under various lighting conditions [48]. For each person, we randomly select 32 images for training and the remaining for testing. Therefore, there will be 1216 training images and 1198 test images. For the experimental settings,  $l_2$  distance and 504-dimension random projection features [3, 51] are used for similarity measure and classification. We evaluate our approach via different dictionary size 380, 570, 760, 950 and 1216, namely 10, 15, 20, 25 and 32 samples for training per category. For settings of KCRC-LCD, we use the global

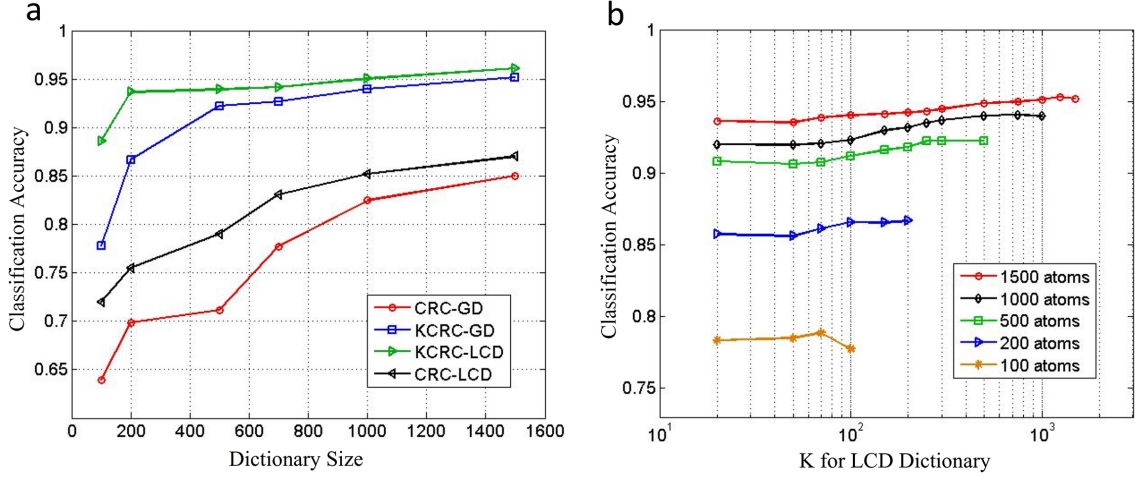


Figure 4. (a) Performance comparison on MNIST under different dictionary size. (b) KCRC-LCD with different size of the global dictionary that generates the LCD under different  $K$  settings. Note that,  $l_2$  distance and  $28 \times 28$  raw pixel features are used for similarity measure and classification.

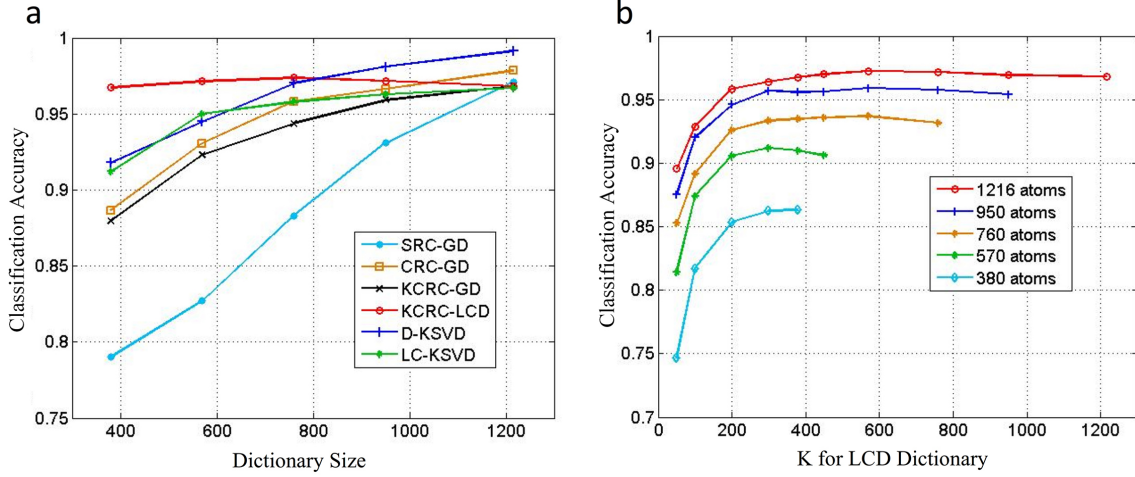


Figure 5. (a) Performance comparison with SRC-GD, CRC-GD, KCRC-GD, KCRC-LCD, LC-KSVD [50, 51] and discriminative K-SVD (D-KSVD) [52] on extended Yale B database under different dictionary size. (b) KCRC-LCD with different size of the global dictionary that generates the LCD under different  $K$  settings. Note that,  $l_2$  distance and the random projection features are used for similarity measure and classification.

dictionary of size 1216 (32 training samples per person) to generate the LCD and set  $K$  for LCD as 380, 570, 760, 950 and 1216 for comparison. Experimental results are given in Fig. 5. In Fig. 5(a), KCRC-LCD has better classification accuracy than the other approaches when dictionary size is small. It is mostly because the global dictionary we use to generate LCD is more informative than the small size dictionary, and LCD itself is designed to be adaptive to use the important information for classification. We can also learn that the classification accuracy of KCRC-LCD no longer stands out when dictionary size becomes 1216. When dictionary size comes to 1216,  $K$  for LCD is also equal to 1216, making KCRC-LCD degenerate to KCRC-GD. That is to say, we do not have enough training samples to construct a discriminative LCD. Moreover, when  $K$  becomes larger, the locality of LCD becomes weaker as well, leading to less discrimination power. It is obtained from Fig. 5(b)

that LCD already has enough critical information to proceed the classification when  $K$  reaches 380. Therefore, we can conclude that the performance ceiling for LCD has been reached when  $K = 380$  in this case. Compared to the global dictionary with 1214 training samples, LCD with only 380 atoms can achieve similar or even better classification accuracy.

### 6.3.3. Caltech101

The Caltech101 database [53] contains 9,144 images from 102 classes (101 objects and a background class). We train on 5, 10, 15, 20, 25, 30 samples per category (dictionary size is 510, 1020, 1530, 2040, 2550, 3060 respectively) and test on the rest. Note that, we use the global dictionary of size 3060 (30 training samples per category) to generate the LCD and set  $K$  for LCD as 510, 1020, 1530, 2040, 2550 and 3060 for comparison.  $l_2$  distance and the spatial pyramid features are used for similarity

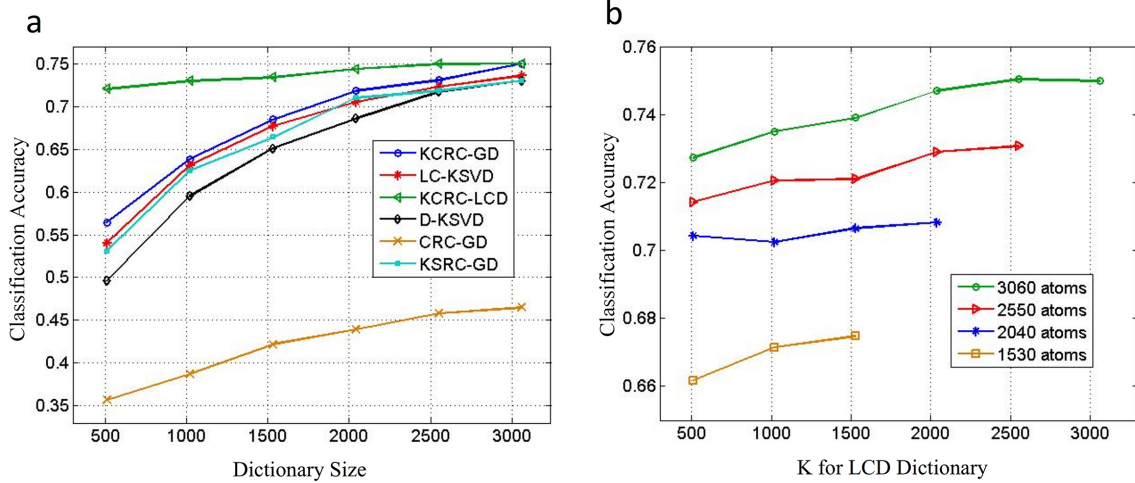


Figure 6. (a) Performance comparison with KCRC-GD, KCRC-LCD, CRC-GD, LC-KSVD [50, 51], D-KSVD [52] and KSRC-GD [14] on Caltech101 under different dictionary size. (b) KCRC-LCD with different size of the global dictionary that generates the LCD under different  $K$  settings. Note that,  $l_2$  distance and the spatial pyramid features are used for similarity measure and classification.

measure and classification. Results in Fig. 6(a) show KCRC-LCD outperforms other competitive approaches in Caltech101 database, especially when dictionary size is small. As we can see in Fig. 6(a), the classification accuracy is improved little when dictionary size is high. The reason is similar to the previous experiment on extended Yale B database. It is due to the lack of extra training samples, or in other word, extra discriminative information, to construct LCD since KCRC-LCD becomes KCRC-GD when  $K$  equals to the 3060. From Fig. 6(b), we can learn that when LCD has obtained the most discriminative and crucial atoms in the dictionary, keeping increasing  $K$  will not help the classification accuracy much. In fact, if we perform the experiment in Fig. 6(b) with smaller  $K$ , it will end up like the curves in Fig. 5(b) where  $K$  obviously has a saturation point for classification accuracy.

#### 6.3.4. Caltech256

The Caltech256 database [54] contains 30607 images of 256 categories, each category with more than 80 images. It is a very difficult visual categorization database due to the large variations in object background, pose and size. We experiment KCRC-LCD on 5, 15 and 30 training samples per category (dictionary size is 1280, 3840 and 7680 respectively). For the settings of KCRC-LCD, use the global dictionary of size 7680 (30 training samples per category) to generate the LCD and set  $K$  for LCD as 1280, 3840 and 7680 for comparison.  $l_2$  distance and 504-dimension Eigenface features are used for similarity measure and classification. Results in Fig. 7 shows when dictionary size is 7680, KCRC-GD, KCRC-LCD and locality constrained K-SVD (LC-KSVD) have similar classification accuracy. While LC-KSVD is slightly better, it can be observed that KCRC-LCD performs better with small dictionary size.

#### 6.3.5. 15 Scene Categories

This database contains 15 natural scene categories such as office, kitchen and bedroom, introduced in [55]. Following the

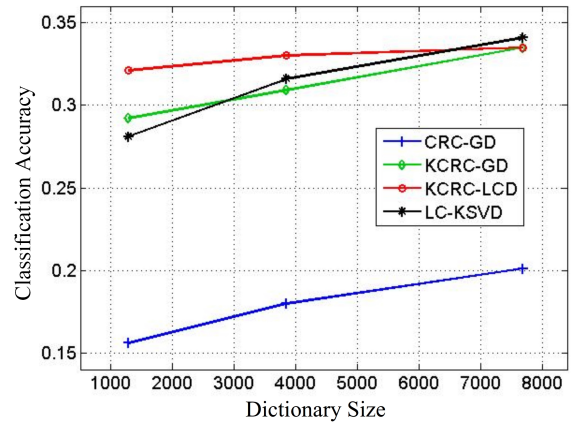


Figure 7. Performance comparison with CRC-GD, KCRC-GD, KCRC-LCD, and LC-KSVD [50, 51] on Caltech256 under different dictionary size. Note that,  $l_2$  distance and Eigenface features are used for similarity measure and classification.

same experimental settings as [50], we randomly select 30 images per category for training and the rest for testing. Note that, we generate the LCD with  $K = 450$  from a global dictionary of size 1500, similar to the training settings of LC-KSVD. Results in Table. 2 and Fig. 8 validate the superiority of KCRC-LCD in scenes.

#### 6.4. Experiments on Running Time

We conduct experiments on running time to evaluate the computational cost of KCRC-LCD. We use public databases including MNIST, extended Yale B, Caltech101 and 15 scene categories to perform our experiments. The detailed experimental settings are given in Table. 3. Note that,  $l_2$  distance is used for similarity measure. For SRC, we use the basis pursuit (BP) algorithm to solve the  $l_1$  minimization problem. Experimental results are shown in Table. 4. Compared to SRC



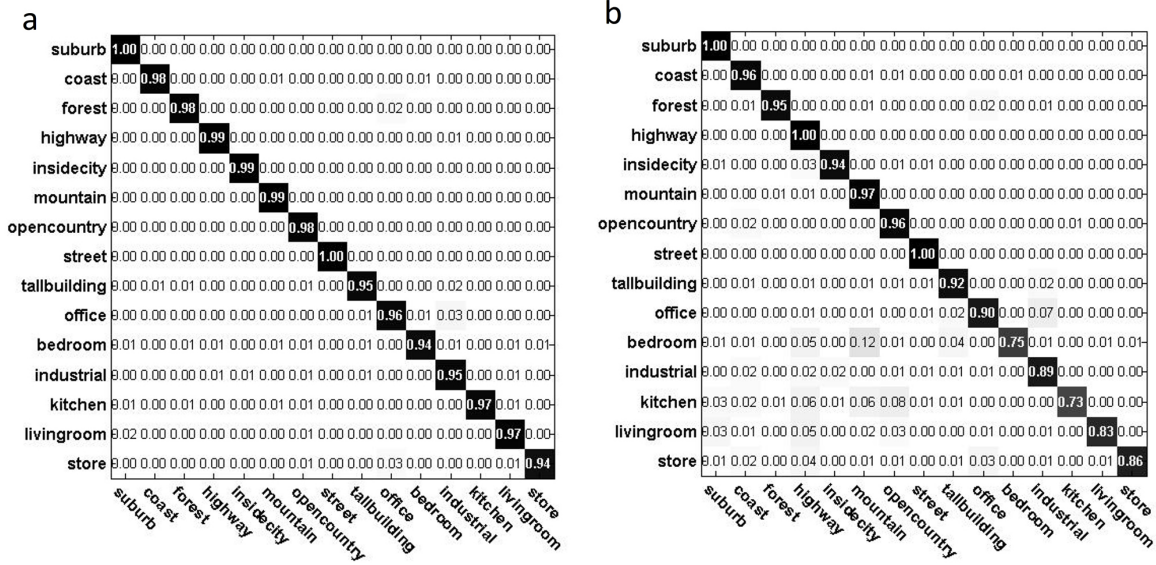


Figure 8. Confusion matrices of (a) KCRC-LCD and (b) CRC-GD on the 15 scene categories database with dictionary size 450. Note that,  $l_2$  distance and the spatial pyramid features are used for similarity measure and classification respectively.

Table 2. Classification results using spatial pyramid features on 15 scene categories database. Both the dictionary size and  $K$  for LCD are set as 450.

Method	Accuracy(%)	Method	Accuracy(%)
KCRC-LCD	98.05	LC-KSVD [50, 51]	92.94
KCRC-GD	97.21	D-KSVD [52]	89.16
CRC-LCD	92.13	KSRC [14] -LCD	96.97
CRC-GD	90.92	KSRC [14] -GD	95.21

approach, KCRC performs much faster due to its  $l_2$  regularization. Constrained with locality, KCRC-LCD performs faster than KCRC-GD and CRC-GD under most circumstances.

### 6.5. Evaluation of The Unified Distance Measurement

Distance metrics are of great importance in the KCRC-LCD, since they grant KCRC-LCD the scalability and discrimination power. Selecting the proper distance metric for the objects can greatly enhance the classification accuracy. Therefore, we validate the superiority of discriminative distance metrics by comparing different distance metrics in USPS, extended Yale B, MNIST and Caltech101 databases. For KCRC-LCD, we use the identity matrix as  $\Psi$ . We use Euclidean distance as baseline for comparison. The USPS database contains 9288 handwritten digits collected from mail envelopes [56]. There are 7291 images for training and 2007 images for testing. This database is fairly difficult since its human error rate is 2.5% [33]. We apply tangent distance to construct the LCD. Specifically, tangents are attained by smoothing each image with a Gaussian kernel of width  $\sigma = 0.7$ . Results are shown in Table. 5. For extended Yale B database, we use the same experimental settings as the previous subsection. We apply the distance metric that is proposed in [57]. In detail, local binary pattern (LBP) histograms are extracted from divided face area and concatenated into a single feature histogram. Then  $\chi^2$  distance is used to measure the similarity of different face histograms. The neighborhood for LBP operator is set as (8, 2) and the window size is  $11 \times 13$ . We term the distance as LBP- $\chi^2$  distance. Results are shown

in Table. 5. The MNIST database [49] of handwritten digits contains 60,000 samples (10 digits) for training and 10,000 for testing. We randomly select 20 samples per digit and construct a global dictionary of 200 size and test on the given 10000 samples.  $K$  for LCD is set as 50 and raw pixel features are used. Results are given in Table. 5. For Caltech101 database [53], we randomly select 30 samples per class and test on the rest (global dictionary size is 3060).  $K$  for LCD is set as 500 and spatial pyramid features are used. Results are given in Table. 5.

Experimental results in Table. 5 show that properly selecting a good distance metric can enhance the discrimination power, and that the performance of different distance metrics can vary significantly as the distance changes (eg., Euclidean distance performs worse than Correlation distance on both USPS and Extended Yale B, but better on MNIST and Caltech101). Such variation is ubiquitous since in general every distance metric only works well under certain situations, and that the characteristics among different datasets are different due to dataset bias. Fortunately, any distance metric can be adopted into our proposed KCRC-LCD framework, showing its flexibility and generalization ability. But while our framework allows such flexibility, the unified distance measurement framework allows to by pass the troublesome process of traversing every single metric to examine its performance. In Table. 5, we can see although the results of unified distance on USPS and extended Yale B databases are not the best, but they are still very close to the optimal one. Basically, one does not need to consider the distance metrics one by one and the unified framework has

Table 3. Experimental settings for running time test.

Database	Feature Dimension	Category	Training Size	Testing Size	$K$ for LCD
MNIST	784	10	500	10000	50
Extended Yale B	504	38	1216	1198	200
Caltech101	3000	102	3060	6084	510
15 Scene Categories	3000	15	1500	2985	200

Table 4. Comparison results of average running time (ms) per classification.

Database	KCRC-LCD	KCRC-GD	CRC-GD	SRC-GD
MNIST	2.2618	3.4723	2.5451	338.69
Extended Yale B	90.434	367.31	200.48	587.31
Caltech101	3016.1	9863.3	2350.9	19147
15 Scene Categories	246.64	2516.7	430.48	5943.8

automatically select the good ones, remeditating the distance metric biases. In addition, the main reason why the unified framework is not the best is because the distance metrics not complementary in this dataset (There are very few cases where other distances can complement the Tangent distance, and including other distances are essentially just poisoning the good results). Also the classification performance is pretty saturated (close to 100%) and the dataset itself not diverse enough. If one uses another distance that is complementary to Tangent distance and LBP- $\chi^2$  distance, we believe the classification accuracy on USPS and extended Yale B databases will be further improved. On MNIST and Caltech101 however, one could see that the combination even brings performance gain over the best single distance.

## 7. Conclusions

We elaborated the KCRC approach in which kernel technique is smoothly combined with CRC. KCRC enhances the discrimination ability of CRC, making the decision boundary more reasonable. Additionally, we present a locality constrained dictionary, of which the locality is exploited to further enhance the classification performance. KCRC and LCD are mathematically linked via distance kernelization. On one hand, LCD not only helps the classifier adaptive and scalable to large databases via pruning the dictionary, but also reduce the dimensionality in kernel space, enhancing both discrimination ability and efficiency. On the other hand, kernel function makes our approach discriminative and robust to more data distribution, i.e., the same direction distribution. Furthermore, the coarse-to-fine classification strategy of KCRC-LCD is similar to the human perception process, which makes the intuition of KCRC-LCD even more appealing.

We conduct comprehensive experiments to show the superiority of KCRC-LCD. Our approach yields very good classification results on various well-known public databases. While achieving high level discrimination ability, efficiency is one of the biggest merits of KCRC-LCD, which is validated in running time test. Moreover, we simulate the representation and construction of KCRC with different dimensionality reduction for kernel space, and further experiment these methods on public database. The simulation results show the discrimination

ability of KCRC. Different distance metrics used in LCD are also compared to support the idea that discriminative distance metric can greatly improve the classification accuracy. We also create a toy data sets to show CRC suffers from data with the same direction distribution while KCRC perfectly overcomes such shortcoming. To sum up, tested by various experiments, KCRC is proven discriminative and efficient when combined with LCD.

Possible future work includes improving the unified similarity measure model and learning the most effective kernel for KCRC-LCD instead of selecting the fixed kernel. It can be predicted that KCRC-LCD will becomes more powerful when combined with kernel learning, or even multiple kernel learning.

## References

- [1] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding, *The Journal of Machine Learning Research* 11 (2010) 19–60.
- [2] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, *Image Processing, IEEE Transactions on* 15 (12) (2006) 3736–3745.
- [3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31 (2) (2009) 210–227.
- [4] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on, IEEE, 2009*, pp. 1794–1801.
- [5] X.-T. Yuan, X. Liu, S. Yan, Visual classification with multitask joint sparse representation, *Image Processing, IEEE Transactions on* 21 (10) (2012) 4349–4360.
- [6] C. Lang, G. Liu, J. Yu, S. Yan, Saliency detection by multitask sparsity pursuit, *Image Processing, IEEE Transactions on* 21 (3) (2012) 1327–1338.
- [7] T. Guha, R. K. Ward, Learning sparse representations for human action recognition, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34 (8) (2012) 1576–1588.
- [8] X. Mei, H. Ling, Robust visual tracking and vehicle classification via sparse representation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33 (11) (2011) 2259–2272.
- [9] R. Rigamonti, M. A. Brown, V. Lepetit, Are sparse representations really relevant for image classification?, in: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011*, pp. 1545–1552.
- [10] Q. Shi, A. Eriksson, A. van den Hengel, C. Shen, Is face recognition really a compressive sensing problem?, in: *Computer Vision and Pattern*

Table 5. Recognition results (%) of different distance metrics on USPS database, MNIST database, extended Yale B database and Caltech101 database.

Distance Metric	USPS	Extended Yale B	MNIST	Caltech101
Euclidean Distance	95.49	96.93	85.31	72.83
Manhattan Distance	94.88	96.26	84.93	70.92
Correlation Distance	95.57	97.01	85.17	72.55
Tangent Distance	97.67	N/A	N/A	N/A
LBP- $\chi^2$ Distance	N/A	98.53	N/A	N/A
Unified Distance	97.17	98.22	86.52	73.05

- Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 553–560.
- [11] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition?, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 471–478.
- [12] B. Schölkopf, A. Smola, K.-R. Müller, Kernel principal component analysis, in: Artificial Neural Networks ICANN'97, Springer, 1997, pp. 583–588.
- [13] C. J. Burges, A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery 2 (2) (1998) 121–167.
- [14] L. Zhang, W.-D. Zhou, P.-C. Chang, J. Liu, Z. Yan, T. Wang, F.-Z. Li, Kernel sparse representation-based classifier, Signal Processing, IEEE Transactions on 60 (4) (2012) 1684–1695.
- [15] W. Liu, L. Lu, H. Li, W. Wang, Y. Zou, A novel kernel collaborative representation approach for image classification, in: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, 2014, pp. 4241–4245.
- [16] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, P. Boyes-Braem, Basic objects in natural categories, Cognitive Psychology 8 (3) (1976) 382–439.
- [17] T. Malisiewicz, A. Gupta, A. A. Efros, Ensemble of exemplar-svms for object detection and beyond, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 89–96.
- [18] S. Gao, I. W.-H. Tsang, L.-T. Chia, Kernel sparse representation for image classification and face recognition, in: Computer Vision–ECCV 2010, Springer, 2010, pp. 1–14.
- [19] J. Yin, Z. Liu, Z. Jin, W. Yang, Kernel sparse representation based classification, Neurocomputing 77 (1) (2012) 120–128.
- [20] W. Liu, Y. Wen, K. Pan, H. Li, Y. Zou, A kernel-based l 2 norm regularized least square algorithm for vehicle logo recognition, in: Digital Signal Processing (DSP), 2014 19th International Conference on, IEEE, 2014, pp. 631–635.
- [21] J. Li, H. Zhang, L. Zhang, Column-generation kernel nonlocal joint collaborative representation for hyperspectral image classification, ISPRS Journal of Photogrammetry and Remote Sensing 94 (2014) 25–36.
- [22] J. Li, H. Zhang, Y. Huang, L. Zhang, Hyperspectral image classification by nonlocal joint collaborative representation with a locally adaptive dictionary, Geoscience and Remote Sensing, IEEE Transactions on 52 (6) (2014) 3707–3719.
- [23] L. Zhang, M. Yang, X. Feng, Y. Ma, D. Zhang, Collaborative representation based classification for face recognition, arXiv Preprint arXiv:1204.2358.
- [24] V. N. Vapnik, Statistical learning theory, Wiley.
- [25] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, K. Mullers, Fisher discriminant analysis with kernels, in: Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop., IEEE, 1999, pp. 41–48.
- [26] F. Orabona, J. Keshet, B. Caputo, The projectron: a bounded kernel-based perceptron, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 720–727.
- [27] W. He, S. Wu, A kernel-based perceptron with dynamic memory, Neural Networks 25 (2012) 106–113.
- [28] O. Dekel, S. Shalev-Shwartz, Y. Singer, The forgetron: A kernel-based perceptron on a budget, SIAM Journal on Computing 37 (5) (2008) 1342–1372.
- [29] C. E. Rasmussen, Gaussian processes for machine learning.
- [30] D. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural computation 16 (12) (2004) 2639–2664.
- [31] A. Y. Ng, M. I. Jordan, Y. Weiss, et al., On spectral clustering: Analysis and an algorithm, Advances in neural information processing systems 2 (2002) 849–856.
- [32] B. Schölkopf, The kernel trick for distances, Advances in neural information processing systems (2001) 301–307.
- [33] H. Zhang, A. C. Berg, M. Maire, J. Malik, Svm-knn: Discriminative nearest neighbor classification for visual category recognition, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, Vol. 2, IEEE, 2006, pp. 2126–2136.
- [34] D. P. Bertsekas, Nonlinear programming, Athena Scientific.
- [35] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, arXiv Preprint arXiv:1009.5055.
- [36] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural computation 15 (6) (2003) 1373–1396.
- [37] P. Zhu, L. Zhang, Q. Hu, S. C. Shiu, Multi-scale patch based collaborative representation for face recognition with margin distribution optimization, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 822–835.
- [38] J. Waqas, Z. Yi, L. Zhang, Collaborative neighbor representation based classification using l2-minimization approach, Pattern Recognition Letters 34 (2) (2013) 201–208.
- [39] Y. Zhou, K. E. Barner, Locality constrained dictionary learning for nonlinear dimensionality reduction, Signal Processing Letters, IEEE 20 (4) (2013) 335–338.
- [40] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, International Journal of Computer Vision 43 (1) (2001) 29–44.
- [41] E. Levina, Statistical issues in texture analysis, Ph.D. thesis, University of California, Berkeley (2002).
- [42] P. Y. Simard, Y. A. LeCun, J. S. Denker, B. Victorri, Transformation invariance in pattern recognition–tangent distance and tangent propagation, in: Neural networks: tricks of the trade, Springer, 2012, pp. 235–269.
- [43] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (4) (2002) 509–522.
- [44] A. C. Berg, J. Malik, Geometric blur for template matching, in: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, Vol. 1, IEEE, 2001, pp. I–607.
- [45] G. Mori, J. Malik, Estimating human body configurations using shape context matching, in: Computer Vision/ECCV 2002, Springer, 2002, pp. 666–680.
- [46] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.
- [47] B. Taati, M. Greenspan, Local shape descriptor selection for object recognition in range data, Computer Vision and Image Understanding 115 (5) (2011) 681–694.
- [48] A. S. Georgiades, P. N. Belhumeur, D. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, Pattern Analysis and Machine Intelligence, IEEE Transactions on 23 (6) (2001) 643–660.
- [49] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
- [50] Z. Jiang, Z. Lin, L. S. Davis, Label consistent k-svd: learning a discriminative dictionary for recognition, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (11) (2013) 2651–2664.
- [51] Z. Jiang, Z. Lin, L. S. Davis, Learning a discriminative dictionary for sparse coding via label consistent k-svd, in: Computer Vision and Pattern

- Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1697–1704.
- [52] Q. Zhang, B. Li, Discriminative k-svd for dictionary learning in face recognition, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 2691–2698.
  - [53] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories, *Computer Vision and Image Understanding* 106 (1) (2007) 59–70.
  - [54] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset.
  - [55] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition (CVPR), 2006 IEEE Conference on, Vol. 2, IEEE, 2006, pp. 2169–2178.
  - [56] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, *Neural computation* 1 (4) (1989) 541–551.
  - [57] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, in: Computer vision-eccv 2004, Springer, 2004, pp. 469–481.